Master 2 Biologie-Informatique/ Bioinformatique 2017-2018

Université Paris Diderot - Paris 7

Etude de la fonction et des mécanismes d'évolution des séquences répétées centromériques chez les Primates

Sarah Kaddah

Tuteur : Loïc Ponger

Structure et Instabilité des Génomes

MNHN - CNRS UMR 7196 / INSERM U1154 - Sorbonne Universités

Remerciements

Merci à Namrod pour toute la partie sur la bibliographie. Retrouvez ses questions FAQ qui ont permis la rédaction de cette partie.
Merci à f-leb, LittleWhite et Metalman pour leurs conseils et la relecture. Merci à ced et jacques_jean pour la correction orthographique et typographique.

Table des matières

1 Introduction

1.1 Les séquences centromériques

-> biblio CENP-A

-> biblio kinetochore

-> info supp sur l'ADN satellite

Le centromère est une structure chromatinienne caractérisé par la présence de CENP-A. Cette protéine, très conservée au cours de l'évolution, est un variant de l'histone H3. Son rôle est de fixer la position du kinétochore par un mécanisme encore peu connu. En effet, le centromère est le site d'assemblage du kinétochore, un ensemble d'ADN et de protéines. Il permet l'attachement du fuseau mitotique pour la ségrégation des chromosomes durant la division cellulaire chez les eucaryotes. Le centromère et les protéines impliquées sont relativement bien conservés. Au contraire, l'ADN sous-jacent est très diversifié et l'organisation varie d'un taxon à l'autre. Cependant, une caractéristique commune est retrouvée chez toute les espèces : de l'ADN centromérique répété en tandem nommé ADN satellite. Ces répétitions sont issues d'événements d'amplification, tels les crossovers inégaux, la conversion de gènes, les cercles roulants ou la transposition de séquences.[Malik and Henikoff, 2002 ; Plohl et al. 2012] Ces séquences représentent 5% du génome. Les répétitions s'étendent de 7pb à 3,2kb avec des séquences de 145-180kb le plus souvent.

1.2 L'ADN α-satellites

-> première mise en évidence des AS

->théorie gradient de l'âge

-> travaux sur le gorille à dev

L'ADN satellite chez les Primates est connu sous le nom d'α-satellite. Ces séquences centromériques répétées en tandem sont riches en AT. -> article sur la première découverte

Des études chez l'Homme propose un modèle évolutif. La répartition des α-satellites suivrait une répartition spécifique selon l'âge des familles. Les familles les plus jeunes s'insèrent au cœur du centromère, repoussant les familles les plus anciennes jusqu'aux regions voisines, appelé péricentromère.

->Article Shepelev

->Est-ce que je peux utiliser du conditionnel ? OU est-ce que cette théorie est confirmée ?

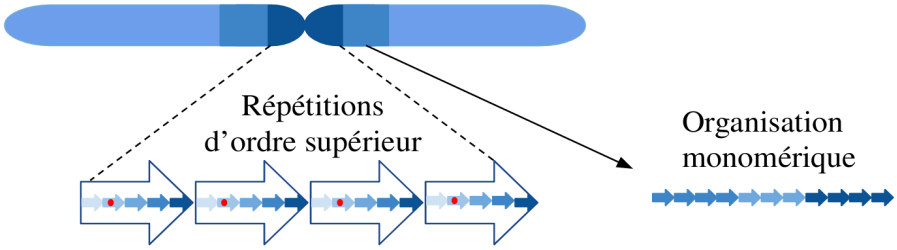

FIGURE 1 – **Organisation spatiale des α-satellites :** Le coeur du centromère (bleu foncé) est organisé en répétition d'ordre supérieur. Le péricentromère(bleu clair) a une organisation monomérique. Un monomère d'une même famille est représenté par une petite flèche de même couleur. Les points rouges représentent les sites de fixation à CENP-B ou pJα.

Un monomère a une longueur de 171pb et il peut être répété des milliers de fois. Les monomères peuvent être répartis en famille selon leur similarité, les séquences ayant un taux d'identité supérieur à 70%. Ces séquences ont soit une organisation monomérique soit une organisation en répétition d'ordre supérieur (Fig. 1). Dans le premier cas, les séquences d'une même famille sont répétés en tandem. Dans le deuxème cas, une suite de monomères appartenant à différentes familles forme une unité, qui elle est répétée en tandem.

Ces séquences peuvent avoir un site de liaison à la protéine centromérique CENP-B un motif spécifique de 17pb. Cette protéine, qui reconnaît et se fixe sur l'ADN, serait présente chez de nombreuses familles de Primates. La protéine pJα, une protéine peu caractérisée, reconnaît un motif qui remplace celui de CENP-B.

Les α-satellites ont essentiellement été étudiées chez l'homme. Modèle évolutif avec les centromères en expansion. Une hypothèse concernant l'âge des séquences découle de ces recherches : les séquences les plus récentes apparaissent au coeur du centromères, déplançant les plus anciennes au péricentromère. D'autres études chez le gorille ont été faites. Le rôle des α-satellites est encore mal connu.

1.3 Le sujet de stage

enchaîne sur l'étude chez les cerco, une autre étude de séquençage haut débit

-travaux précédents limités (expliquer pk). Les méthodes basées aur l'alignement et la phylogénie sont très limitées, le jeu de données étant trop grand. Les méthodes n'étaient pas objectives (quelles méthodes ??). De plus, chez d'autres espèces de Primates, les informations sont trop dispersées et aucune comparaison interespèce n'a été faite.

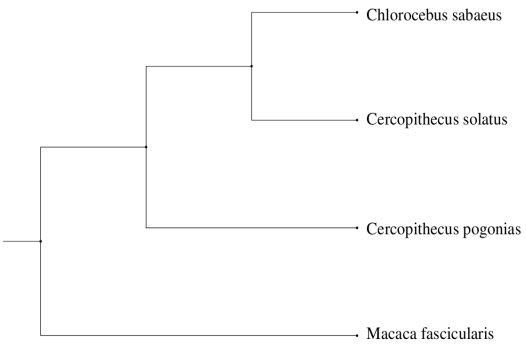

FIGURE 2 – **Arbre phylogénétique des espèces choisies.**

L'équipe d'accueil de mon stage "ADN répété, Chromatine, Evolution" ou ARChE, a récemment développé une approche de séquençage haut débit, ciblée sur les séquences α-satellites chez deux espèces de Cercopithèques. Une autre étude avec un grand nombre de séquences concerne le Gorille [Catacchio] avec l'utilisation de fragments relativement longs.

L'objectif de ce stage est de comprendre la fonction des α-satellites et leur mécanisme d'évolution. Je vais choisir plusieurs espèces de Primates. Je vais utiliser une méthode de classification automatisée améliorée du laboratoire [Florence Jornod] pour classer les séquences en familles. Ce programme permet de traiter des centaines de milliers de séquences sans quelque soit le nombre de séquences ou la taille des familles. Je vais dans un premier temps appliquer cette technologies au données issues de ce séquençage. Ensuite, je vais étudier d'autres espèces. Puisque toutes les espèces sont étudiées par la même méthode, une comparaison inter espèce est envisageable.

2 Matériel et méthode

2.1 Choix des espèces

Les critères de sélection dépendent de la disponibilité des séquences de qualité. Deux espèces du laboratoire sont choisies, les *Cercopithèques solatus* et *pogonias*, et deux espèces proches, le *Macaca fascicularis* et le *Chlorocebus sabaeus*.

2.2 Méthode de classification

Cette méthode [2] répartit des séquences α-satellites en familles selon la similarité. La classification est hiérarchique dichotomique. Dans un premier temps, les séquences sont séparées en fonction de la fréquence des 5-mers qui composent les séquences, d'après les études sur les *Cer-*

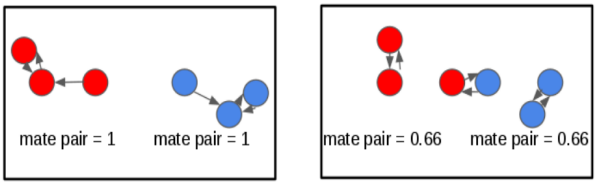

FIGURE 3 – **Représentation du matepair :** Dans l'exemple de droite, les séquences "rouges" et les séquences "bleues" forment des groupent distincts. Leur pairmate est de 1. Dans l'exemple de droite, une séquence rouge est plus proche d'une séquence bleue, diminuant la valeur du pairmate à 0,66.

copithèques.

La classification est suivie d'une double validation des sous-groupes. D'une part la taille du sous-groupe est vérifiée. La taille minimale d'une famille est fixée à 100. Si un groupe atteint 100 séquences, il n'est pas redivisé. D'autre part les deux groupes doivent être distincts. Pour cela le *matepair*, la proportion de monomères ayant son plus proche voisin dans le même groupe, est évalué. Des valeurs *matepairs* élevées indiquent des sous-groupes bien homogènes et séparés validant la classification tandis qu'un seuil *matepair* plus faible entraîne plus de classes. Si les *matepairs* sont au-dessus d'un certain seuil, les deux sous-groupes sont ajoutés séparément à la file pour être potentiellement redivisés ultérieurement.En revanche, si au moins une des valeurs de *matepair* est au dessous de ce seuil, les sous-groupes sont considérés comme formant un seul groupe et le groupe initial est sauvegardé comme une famille unique.

La séparation des séquences se fait de façon itérative en boucle. Chaque tour implique une analyse en composante principale (ACP) et une classification hiérarchique. Si le jeu de données dépasse 110 000 séquences, le calcul des distances devient pesant. Une analyse discriminante linéaire (LDA) entre alors en jeu, avec un échantillon de taille 100 000. L'ACP est faite sur la table des 5-mers pour réduire les dimensions du jeu de données en minimisant la perte d'information et obtenir des variables indépendantes utilisables pour la LDA. Le nombre de composantes est fixé à 1024. Ensuite des distances euclidiennes sont calculées entre toutes les paires de séquences dans l'espace défini par les M premières composantes de l'ACP. Puis la méthode de classification hiérarchique de Ward forme des classes de façon à minimiser l'inertie interclasse. Cette méthode d'apprentissage utilise un sous-jeu de données formé par des séquences tirées aléatoirement. Le modèle construit est appliqué sur toutes les séquences.

2.3 Alignement, consensus et phylogénie

L'alignement des séquence est fait avec muscle [1]. La phylogénie est reconstruite avec Seaview [3] utilisant le modèle évolutif de Kimura à deux paramètre (K2P). L'arbre est construit avec l'algorithme de Neighbor Joining (BioNJ) [4]. Les consensus sont obtenus avec des scripts développés par l'équipe. Les motifs CENP-B (TTCGTTGGAA[AG]CGGGA), PJα (TTCCTTTT[CT]CACC[AG]TAG) et pKβ (CTATAGGGCCAAAGGAA) ont été identifiés avec le logiciel fuzznuc (package EMBOSS) [5] et en autorisant 2 différences au maximum par rapport au consensus.

3 Résultat

3.1 Caractérisaton des familles dans plusieurs espèces

3.1.1 Identification des familles

Les espèce *C. solatus* et *C. pogonias* sont analysés dans un premier temps pour comparer les résultats de la classification automatisée avec les résultats expérimentaux du laboratoire. Il existe 6 familles α-satellites chez les *Cercopithèques*. Ces deux espèces partagent deux grandes familles monomériques, C1 et C2, de l'ordre de plusieurs milliers de séquences. Elles partagent également le dimère C3-C4. *C. pogonias* possède les familles supplémentaires C5 et C6.
La classification automatisée donne des familles de tailles variables allant de deux à des dizaines de milliers de séquences. Seules les familles ayant plus de 100 séquences, appelées "grandes familles", sont conservées pour l'analyse des familles. Les "petites familles" sont prises en compte en terme de pourcentage de séquences qui ne figurent pas dans l'analyse. Les séquences α-satellites chez *solatus* sont réparties en 564 familles, dont 12 grandes familles. Les séquences qui ne sont pas retenues représentent 3,97% du jeu de données. Chez pogonias, le nombre total de familles est de 132, avec 13 grandes familles et 1,29% du jeu de données qui ne figure pas dans les analyses.
Le nombre de grandes familles est relativement proche entre ces deux espèces, mais diffèrent significativement des résultats expérimentaux. Parmi ces dizaines de familles, chez *C. solatus* 11 familles forment la famille C2, une familles forme la famille C1 et les familles C3 et C4 sont retrouvées dans des petites familles d'environ 80 séquences chacune. Toutes les familles sont retrouvées chez *C. pogonias* sauf la famille C6. La famille C1 est composée de 10 familles, les familles C2, C3 et C5 sont composé d'une famille respectivement, et C4 correspond a une petite famille. Ces résultats contredisent les résultats expérimentaux. La famille C2 et censée être divisée en plusieurs groupes par ses séquences qui divergent plus que dans la famille 1. La classification expérimentale est une

méthode visuelle basée sur l'ACP sur des 5-mers. Un point noir représente un monomère. Deux groupes distincts regroupent les familles C1 et C2 chez solatus et 4 groupes distincts sont retrouvés chez pogonias formant les familles C1, C2, C5 et C6. Sur cette représentation sont superposées les familles de la classification automatisée en couleur.
Le *C. sabaeus* a 338 familles au total, dont 44 grandes familles, et 10,89% du jeu de données qui n'est pas pris en compte. Le *M. fascicularis* a respectivement 709 et 998 familles, dont 42 et 81 grandes familles, et 14,56% et 5,05% du jeu de données qui n'est pas pris en compte. Ces espèces ont beaucoup plus de grandes familles que les *Cercopithèques*.

3.1.2 Motifs CENP-B, pJα et pKβ

3.1.3 Similarité entre familles

3.2 Comparaison inter-espèce et mécanismes d'évolution

3.2.1 Résultats du laboratoire

Le laboratoire a effectué un séquençage sur les *C. solatus* et *pogonias*. Les séquences α-satellites obtenues sont divisées en nucléotides de 5-mers, puis triées par Analyse de Classification Hiérarchique (HCA) et par Analyse Discriminante Linéaire (LDA). Chez le *C. solatus*[article 1 de Laurianne] deux grandes familles monomériques (C1 et C2) et un HOR d'ordre 2, composé des familles C3 et C4, sont identifiés. Chez le *C. pogonias*[article2 de Laurianne], deux familles supplémentaires (C5 et C6) ont été détectées. Ces *Cercopithèques* ne disposent pas du site de fixation pour la protéine CENP-B. Au contraire, le site de fixation pour la protéine pJα est présent dans 85% des α-satellites en moyenne, excepté la famille C3 qui n'en possède pas.

3.2.2 Comparaison des familles

Le programme identifie 13 et 14 familles chez le *C. solatus* et *pogonias* respectivement. Seules les familles de plus de 100 séquences sont conservées. Cette sélection implique 6,99% et 1.63% en perte d'information pour ces deux espèces. Les séquences restantes, n'ayant pas été classées dans une famille assez grande ou étant peut-être des séquences atypiques, ne sont pas prises en compte dans l'analyse.

Le *C.solatus* a deux grandes familles de 82911 et 9216 monomères, une classe intermédiaire de 1519,1267 et 898 séquences. Le *C.pogonias* a 3 grandes familles d'α-satellites qui ont 94594, 8319 et 5202 séquences, deux familles intermédiaires de 998 et 664 monomères. Les familles restantes ont quelques centaines de séquences et sont très petites comparées au familles citées ci-dessus.

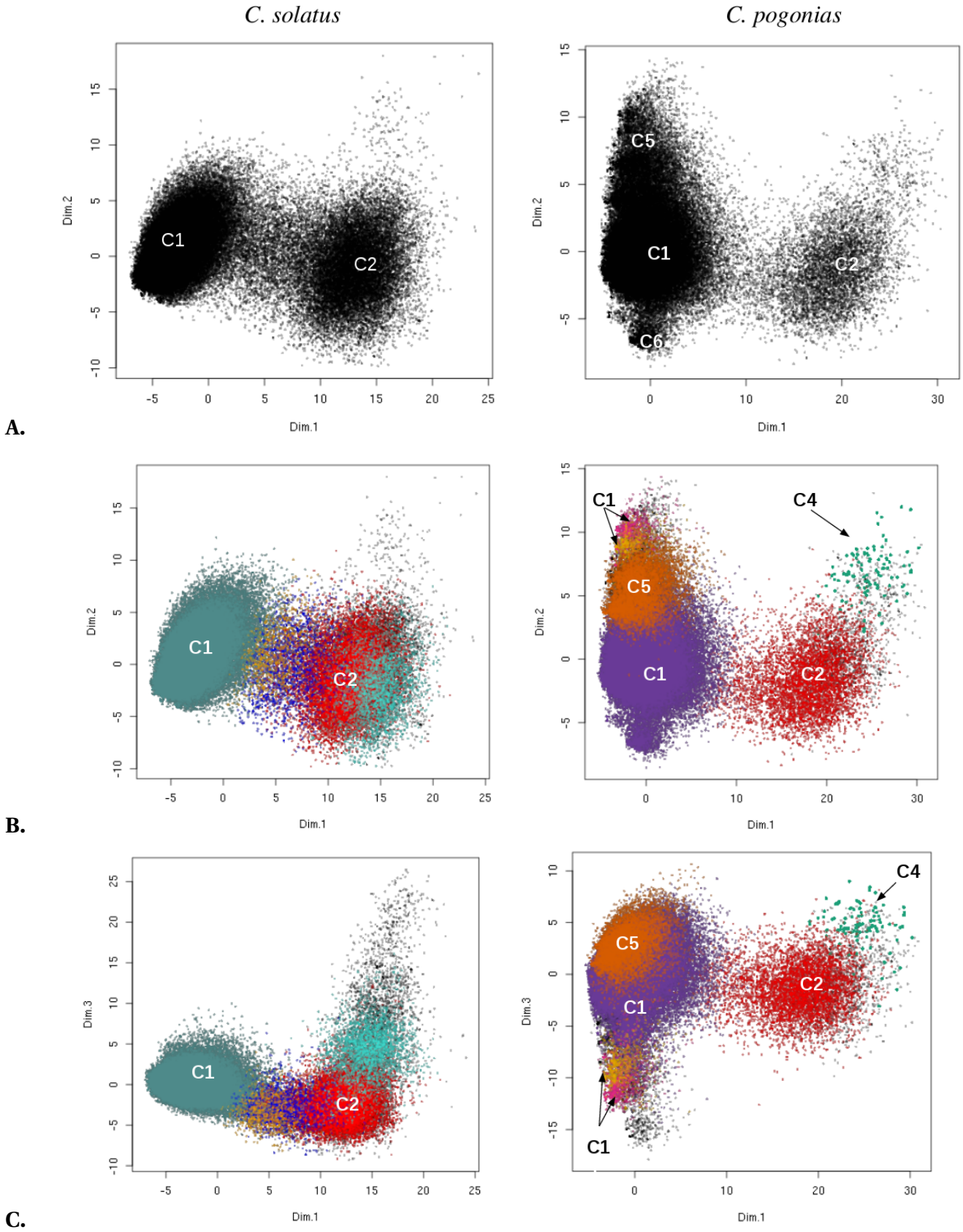

FIGURE 4 – **Représentation des ACP des 5-mers chez *C. solatus* et *C. pogonias* :**Les familles expérimentales C1 à C5 sont indiquées sur les graphiques. Chez solatus, les familles issues de la classification automatisée sont C1 en darkslategray4 et C2 en rouge, blue, turquoise, orange. Chez pogonias, C1 est en violet, jaune et rose ; C2 est en rouge ; C4 est en vert ; C5 est en orange.**A.** Classification expérimentale. Un monomère est représenté par un point noir. **B.** Représentation des composantes 1 et 2 de l'ACP. **C.** Représentation des composantes 1 et 3 de l'ACP..

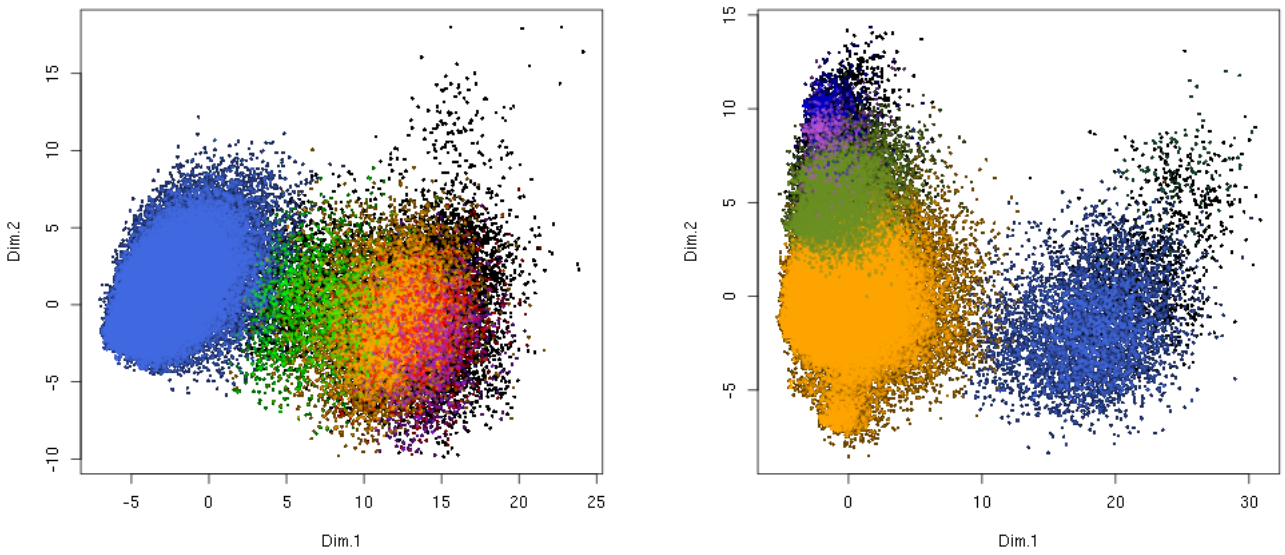

FIGURE 5 – ACP des 5-mers chez *C.solatus* (à gauche) et *C.pogonias* (à droite) : Les familles communes sont C1 (orange) et C2 (royalblue). C3 (seagreen) et C5 (olivedrab) sont visibles seulement chez pogonias. Les familles qui divisent C1 sont la 2 (green), 73 (red) et 177 (purple)et chez pogonias la 11 (mediumblue) et 51 (mediumorchid).

Une ACP permet de visualiser et de comparer les familles (Figure 2).

La famille C1 est divisée en plusieurs familles pour les deux espèces.La famille C2 péricentromérique est retrouvée chez les deux espèces et forme un complexe homogène. Les dimères sont une famille d'une centaine de monomères. La famille C3 est retrouvée chez le *C.pogonias* seulement, bien qu'elle n'ait pas été étudiée dans les études du *C.pogonias*. La famille C4 n'est pas détectée. Cette famille a probablement été divisée en petites classes. Elles ne sont donc pas détectées. La famille C5 est retrouvée. C6 et C1 ont été classé comme étant une même famille.

4 Discussion

5 Conclusion

Références

[1] Robert C. Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, vol. 32(no. 5) :pp. 1792–1797, 2004.

[2] Florence Jornod. Développement d'une méthode de classification pour les séquences répétées centromériques de primates. Master's thesis, Université Paris Diderot - Paris 7, 2016-2017.

[3] Olivier Gascuel Manolo Gouy, Stéphane Guindon. Seaview version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27(2) :221–224, 2009.

[4] Masatoshi Nei Naruya Saitou. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4) :406–425, 1987.

[5] Alan Bleasby Peter Rice, Ian Longden. Emboss : the european molecular biology open software suite. 2000.

Résumé

Votre résumé commence ici... ...

Abstract

Abstract begins here... ...